ESTIMATION OF RELATIVE TRANSFER FUNCTION IN THE PRESENCE OF STATIONARY NOISE BASED ON SEGMENTAL POWER SPECTRAL DENSITY MATRIX SUBTRACTION

Xiaofei Li[1], *Laurent Girin*[1,2], *Radu Horaud*[1]

[1]INRIA Grenoble Rhône-Alpes
[2]GIPSA-Lab & Univ. Grenoble Alpes

Sharon Gannot

Faculty of Engineering
Bar-Ilan University

ABSTRACT

This paper addresses the problem of relative transfer function (RTF) estimation in the presence of stationary noise. We propose an RTF identification method based on segmental power spectral density (PSD) matrix subtraction. First multiple channel microphone signals are divided into segments corresponding to speech-plus-noise activity and noise-only. Then, the subtraction of two segmental PSD matrices leads to an almost noise-free PSD matrix by reducing the stationary noise component and preserving non-stationary speech component. This noise-free PSD matrix is used for single speaker RTF identification by eigenvalue decomposition. Experiments are performed in the context of sound source localization to evaluate the efficiency of the proposed method.

Index Terms— microphone array, relative transfer function, stationary noise.

1. INTRODUCTION

The relative transfer function (RTF) is the acoustic transfer function (ATF) ratio between a given microphone and a reference microphone [1]. It is widely used in array processing, beamforming, sound source separation, and sound source localization [2]. This paper addresses the estimation of a multichannel RTF corresponding to a speech source of interest in the presence of stationary noise. The authors of [1] have proposed to estimate the RTF based on the stationarity of noise signal and the non-stationarity of desired signal, whose exploitation for system identification originated in [3]. Several successive frames are clustered into segments, the cross power spectral density (PSD) between two channels of all the segments are combined to form an overdetermined set of equations, and a least squares solution leads to an unbiased cross-channel RTF estimate. Such an approach has the limitation that a significant amount of noise segments is included. An RTF identification method based on speech-presence probability and spectral subtraction was proposed in [4]. This method takes into account only the segments that have large speech presence probability, at each frequency. Its performance depends on the accuracy of noise PSD estimation. Both types of methods estimate an RTF for each channel pair separately and are suited only for the case of a single desired signal. Other approaches include independent component analysis, e.g., [5] which is applicable in underdetermined scenarios, and [6] that uses sparsity.

In this paper, we propose an RTF identification method based on segmental PSD matrix subtraction. Similarly, successive frames are clustered into segments, then the PSD matrix of each segment is computed. We classify segments into two classes that have high speech power and low speech power respectively. Since the noise signal is assumed stationary, subtracting a PSD matrix with low speech power from a PSD matrix with high speech power leads to a quasi noise-free segmental PSD matrix. Averaging all the segmental noise-free PSD matrices results in a matrix that compacts speech power spectra. Inspired by [7] we then calculate the principal eigenvector of this matrix to provide a good and robust estimation of the RTF in the case of a single speaker. To process the segment classification, we use the minimum and maximum statistics of the noise signal. Minimum statistics were already proposed in [8] and in [9] for noise PSD estimation, for the case where the increment between segments is a single frame. We exploit here an analytical minimum and maximum statistics and use an equivalent sequence length that is suitable for any arbitrary increment between segments.

The remainder of this paper is organized as follows. Section 2 formulates the problem and Section 3 describes the proposed method. Experiments are presented in Section 4 and Section 5 draws some conclusions.

2. PROBLEM FORMULATION

Let us consider a desired directional sound source $s_s(t)$ and a directional[1] noise source $s_i(t)$ recorded by a M-microphone array. The m-th microphone signal in the time domain is given by

$$x_m(t) = h_{s,m}(t) * s_s(t) + h_{i,m}(t) * s_i(t), \qquad (1)$$

This research has received funding from the EU-FP7 STREP project EARS (#609465).

[1]The problem is formulated for a directional noise but the approach is extendable to diffuse stationary noise.

where $h_{s,m}(t)$ and $h_{i,m}(t)$ are the respective source-to-microphone acoustic impulse responses and $*$ denotes convolution. Applying the short-time Fourier transform (STFT), (1) is approximated in the time-frequency (TF) domain as

$$\mathbf{x}(l,\omega) = \mathbf{h}_s(\omega)s_s(l,\omega) + \mathbf{h}_i(\omega)s_i(l,\omega), \quad (2)$$

where $l = 1 \dots L$ is the frame index, $\omega = 1 \dots \Omega$ is the frequency index, $\mathbf{x}(l,\omega)$ is M-channel STFT spectrum of microphone signals, $s_s(l,\omega)$ and $s_i(l,\omega)$ are the STFT spectra of the two sources, and $\mathbf{h}_s(\omega)$ and $\mathbf{h}_i(\omega)$ are the (time-invariant) M-channel acoustic transfer functions from the desired and noise sources to the microphone array, respectively. The problem that we address is the estimation of the RTF of the desired source, i.e., a relative, arbitrarily normalized version of the $\mathbf{h}_s(\omega)$ vector for each frequency ω.

3. THE PROPOSED METHOD

3.1. Segmental PSD Matrix Subtraction

In order to identify the RTF of the desired source, we propose the following PSD matrix subtraction method. First let us define a segment as the concatenation of W successive frames:

$$\mathbf{X}_{l'}(\omega) = [\mathbf{x}((l'-1)R+1,\omega), \dots, \mathbf{x}((l'-1)R+W,\omega)], \quad (3)$$

where R is the segment increment and $l' = 1 \dots L'$ is the segment index. Assuming that $s_s(t)$ and $s_i(t)$ are uncorrelated, the PSD matrix of a segment is given by[2]

$$\begin{aligned}\mathbf{\Phi}_{l'}(\omega) &= \mathbf{X}_{l'}(\omega)\mathbf{X}_{l'}^H(\omega) \\ &\approx \mathbf{h}_s(\omega)\mathbf{h}_s^H(\omega)\Phi_{l'}^s(\omega) + \mathbf{h}_i(\omega)\mathbf{h}_i^H(\omega)\Phi_{l'}^i(\omega), \quad (4)\end{aligned}$$

where $\Phi_{l'}^s(\omega) = \sum_{l=(l'-1)R+1}^{(l'-1)R+W} |s_s(l,\omega)|^2$ is the power summation of the desired source signal in the l'-th segment, and similarly, $\Phi_{l'}^i(\omega)$ is the power summation of the noise signal. This PSD matrix comprises two matrices spanned by the ATF of desired source and noise source respectively, and take their power as weight. $\Phi_{l'}^i(\omega)$ is the smoothed power spectrum using W frames, and has a small variance due to $s_i(t)$ stationarity. Conversely, the fluctuations of $\Phi_{l'}^s(\omega)$ are large because of the non-stationarity and sparsity of speech signals. Therefore, if we calculate the difference between segmental PSD matrices

$$\begin{aligned}\mathbf{\Phi}_{l_1'}(\omega) - \mathbf{\Phi}_{l_2'}(\omega) &= \mathbf{h}_s(\omega)\mathbf{h}_s^H(\omega)(\Phi_{l_1'}^s(\omega) - \Phi_{l_2'}^s(\omega)) \\ &+ \mathbf{h}_i(\omega)\mathbf{h}_i^H(\omega)(\Phi_{l_1'}^i(\omega) - \Phi_{l_2'}^i(\omega)), \quad (5)\end{aligned}$$

the difference of power weights is likely to be much smaller (in absolute value) for the noise signal than for the speech signal, i.e., $|\Phi_{l_1'}^i(\omega) - \Phi_{l_2'}^i(\omega)| \ll |\Phi_{l_1'}^s(\omega) - \Phi_{l_2'}^s(\omega)|$. Consequently, the PSD difference matrix (5) will match the matrix spanned by $\mathbf{h}_s(\omega)$ well.

Ensuring that the difference $|\Phi_{l_1'}^s(\omega) - \Phi_{l_2'}^s(\omega)|$ is large can be done by classifying segments into two classes $\mathbf{l}_1$ and $\mathbf{l}_2$, which have high speech power and low speech power, respectively. This is done in the next section using the minimum and maximum statistics of noise spectrum. Then, (5) is applied for each segment $l_1' \in \mathbf{l}_1$, taking the corresponding segment l_2' (denoted as $l_2'(l_1')$) as its nearest segment in $\mathbf{l}_2$, since in practice, the closer the two segments are, the smaller is the difference of their noise PSD and transfer function.

3.2. Segment Classification

From (4), the trace of the PSD matrix $\mathbf{\Phi}_{l'}(\omega)$ is

$$\xi_{l'}(\omega) = \mathbf{h}_s^H(\omega)\mathbf{h}_s(\omega)\Phi_{l'}^s(\omega) + \mathbf{h}_i^H(\omega)\mathbf{h}_i(\omega)\Phi_{l'}^i(\omega). \quad (6)$$

It is the summation of the power of the *image* desired speech signal and noise signal, i.e., those signals as recorded at the microphones. The minimum statistics approach has been proposed in [9] where the minimum value of (6), multiplied by a bias correction factor, is used as the estimation of noise PSD. Also, in [9] successive smoothed periodograms are processed recursively, or equivalently the increment between segments is a single frame. In this paper we propose to use classification (for two classes $\mathbf{l}_1$ and $\mathbf{l}_2$) thresholds defined from ratios between maximum and minimum statistics. Moreover, our segment increment R in (4) can be an arbitrary integer value from 1 to W, and we introduce an equivalent sequence length for analyzing the minimum and maximum statistics of noise PSD. Finally, we classify the segments by using the minimum controlled maximum border.

Formally, the power of the image noise signal is

$$\xi_{l'}^i(\omega) = \mathbf{h}_i^H(\omega)\mathbf{h}_i(\omega)\sum_{l=(l'-1)R+1}^{(l'-1)R+W} |s_i(l,\omega)|^2. \quad (7)$$

For a stationary signal, the probability density function (pdf) of periodogram bin $|s_i(l,\omega)|^2$ obeys the exponential distribution with a variance equal to the signal PSD, i.e. $\sigma_i^2(\omega) = E\{|s_i(l,\omega)|^2\}$ [9]. Therefore, assuming that $|s_i(l,\omega)|^2$ at different frames are i.i.d. random variables, $\xi_{l'}^i(\omega)$ obeys the Erlang distribution [11]:

$$f(y)|_{y=\xi_{l'}^i(\omega)} = \frac{y^{k-1}e^{-\frac{y}{\mu}}}{\mu^k(k-1)!} \quad y \geq 0, \quad (8)$$

with scale parameter $\mu = \mathbf{h}_i^H(\omega)\mathbf{h}^i(\omega)\sigma_i^2(\omega)$ and shape parameter $k = W$. We are interested in characterizing and estimating the ratio between the maximum and minimum statistics. Since the maximum and minimum statistics are both linearly proportional to μ [9], without loss of generality we assume $\mu = 1$. Consequently the mean value of (8) equals W. If there is no overlap between two adjacent segments, namely

[2] Assuming the decorrelation and stationarity of both source signals, the exact equality holds for the corresponding theoretical PSDs defined as expectations. Here we have an approximation since we work with measured STFT spectra and a non-stationary speech signal. This approximation is assumed to be quite good in practice.

$R = W$, the segmental power sequence $\xi^i_{l'}(\omega),\ l' = 1 \dots L'$ is an independent random sequence. The pdfs of the minimum and maximum of these L' independent variables are [8]:

$$f_{min}(\xi) = L' \cdot (1 - F(\xi))^{L'-1} \cdot f(\xi), \quad (9)$$

$$f_{max}(\xi) = L' \cdot (F(\xi))^{L'-1} \cdot f(\xi), \quad (10)$$

where $F(\cdot)$ denotes the cumulative distribution function (cdf) associated with the pdf (8). Conversely, if $R < W$, $\xi^i_{l'}(\omega)$ is a correlated sequence, and the correlation coefficient is linearly proportional to the overlap. In order to make (9) valid for the correlated sequence, simulations over a large dataset show that an approximate equivalent sequence length

$$\tilde{L'} = \frac{L'R}{W} \cdot \left(1 + \log\left(\frac{W}{R}\right)\right) \quad (11)$$

can replace L' in (9). Then, the expectation of the minimum can be approximately computed as

$$\bar{\xi}_{min} \approx \sum\nolimits_{\xi_i} \xi_i \cdot f_{min}(\xi_i) / \sum\nolimits_{\xi_i} f_{min}(\xi_i), \quad (12)$$

where $\xi_i = \{0, 0.1W, 0.2W, \dots, 3W\}$ is a grid used to approximate the integral operation, which covers well the support of Erlang distribution with shape W and scale 1. Similarly, the cdf of the maximum can be estimated as $F_{max}(\xi) \approx \sum_{\xi_i} f_{max}(\xi_i)$. Finally, we define two classification threshold factors that are two specific values of the maximum to minimum ratios, namely $r_1 = \frac{\xi_{F_{max}(\xi)=0.95}}{\bar{\xi}_{min}}$ and $r_2 = \frac{\xi_{F_{max}(\xi)=0.5}}{\bar{\xi}_{min}}$. The two classes $\mathbf{l}_1$ and $\mathbf{l}_2$ are then obtained as

$$\mathbf{l}_1 = \{l' \mid \xi_{l'}(\omega) > r_1 \cdot \min\{\xi_{l'}(\omega)\}\}, \quad (13)$$

$$\mathbf{l}_2 = \{l' \mid \xi_{l'}(\omega) < r_2 \cdot \min\{\xi_{l'}(\omega)\}\}. \quad (14)$$

These two thresholds are set differently to ensure that the segments in $\mathbf{l}_1$ involve considerable speech power and the segments in $\mathbf{l}_2$ involve negligible speech power. The speech power for the other segments are probabilistically uncertain, making them unsuitable for either $\mathbf{l}_1$ or $\mathbf{l}_2$.

As an illustration of (11), Fig. 1 shows the cdf for $W = 18$. The empirical curves are simulated using white Gaussian noise (WGN) as the stationary noise signal, and the analytical curves are computed using the equivalent sequence length in (11). Three groups of curves are shown for the minimum cdf and maximum cdf, respectively, and $\tilde{L'}$ is fixed for each group. For example, if $\tilde{L'} = 10$, the corresponding segment numbers are 46, 12 and 10, for the three segment increments $R = 1, 9, 18$, respectively. This shows that the equivalent sequence length in (11) is accurate for minimum and maximum statistics.

3.3. RTF Estimation

Finally, in order to obtain a robust RTF estimation, we calculate the *global noise-free PSD matrix* as the sum of noise-free PSD matrices for all segments in $\mathbf{l}_1$, i.e., $\hat{\mathbf{\Phi}}(\omega) = \sum_{l'_1 \in \mathbf{l}_1} (\mathbf{\Phi}_{l'_1}(\omega) - \mathbf{\Phi}_{l'_2(l'_1)}(\omega))$. The principal eigenvector $\mathbf{u}_1(\omega)$ of $\hat{\mathbf{\Phi}}(\omega)$ is then a good estimation of the unit-norm RTF vector corresponding to $\mathbf{h}_s(\omega)$ [7].

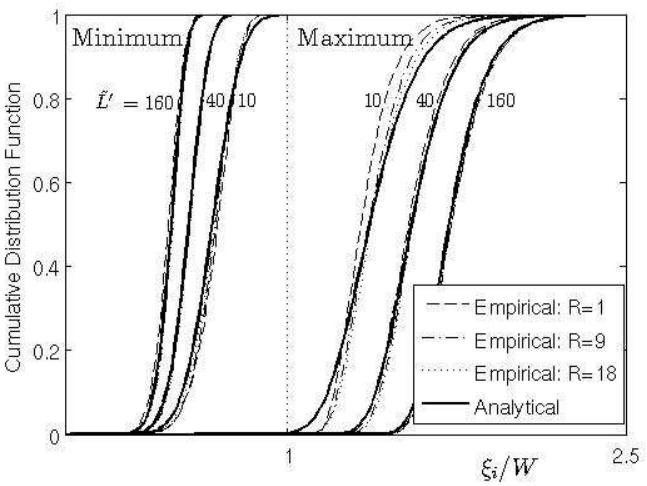

Fig. 1: Cumulative distribution function for $W = 18$ (see text for explanations).

4. EXPERIMENTS: APPLICATION TO SOUND SOURCE LOCALIZATION

4.1. Principle

Evaluation of the RTF estimation technique by comparing true RTF values with estimated RTF values is delicate since the effect of estimation error on practical systems using the RTF is difficult to characterize. Therefore, several authors have preferred to directly measure the performances of systems exploiting RTFs, e.g., in a beamforming framework [1][10][4]. In this paper, we apply this principle in the context of sound source localization (SSL). In the single source-of-interest case, SSL provides a relevant framework for RTF estimation assessment since, in a given environment, there exists a smooth mapping between RTF values and source location [2]. In the present study, we adopt a basic supervised "look-up table" approach: We have available a dictionary $D_{\mathbf{h},\mathbf{p}}$ of K pairs $\{\mathbf{h}_k, \mathbf{p}_k\}_{k=1}^K$, where $\mathbf{h}_k$ is an RTF *feature vector* of a sound source and $\mathbf{p}_k$ is the corresponding source direction vector, for a given room and given microphones position in the room. An RTF feature vector is simply the concatenation of RTF vectors at all frequency bins, $\mathbf{h} = [\mathbf{h}(1)^T, \dots, \mathbf{h}(\Omega)^T]^T$. This dictionary was obtained from noise-free single-source controlled recordings (see next section). Then, for any new RTF feature vector $\tilde{\mathbf{h}}$ extracted from a source + noise test recording, the direction of the source is estimated by selecting the closest vector in $D_{\mathbf{h},\mathbf{p}}$:

$$\hat{\mathbf{p}} = \mathbf{p}_{k_0} \quad \text{with} \quad k_0 = \underset{k \in [1,K]}{\operatorname{argmin}} \parallel \tilde{\mathbf{h}} - \mathbf{h}_k \parallel . \quad (15)$$

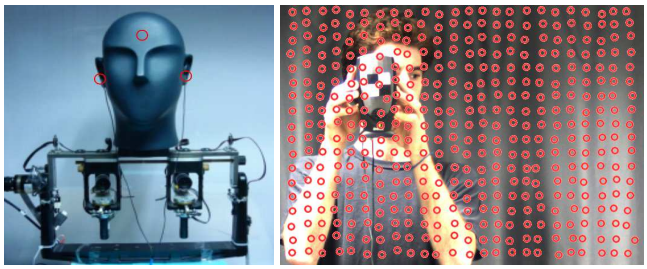

Fig. 2: Audio-Visual dataset. Left: Dummy head, microphones (red circles) and cameras. Only one camera is used here. Right: Camera view with training directions.

4.2. The Dataset

We used an acoustic dummy head equipped with a 4-microphone array (left/right and front/back) and a camera, e.g., Fig. 2(left). The audio-visual data acquisition method is detailed in [12]. To summarize, the training data consists of 1 s white-noise signals emitted by a loud-speaker from 432 source directions spanning a field of view of $24^\circ \times 18^\circ$ (azimuth and elevation). The ground-truth source direction are obtained by localizing the loud-speaker's visual marker in the image provided by the camera. The image resolution is of 640×480 pixels, and 1° of azimuth/elevation corresponds to 23.3 horizontal/vertical pixels. Fig. 1(right) shows the camera field-of-view and the 432 training directions. The test data are 108 speech signals from the TIMIT dataset [13] emitted by the same loud-speaker from 108 directions in the camera field-of-view. All source positions are located in a plane at approximately 2.5 meters away from the dummy head. White Gaussian noise (WGN) and babble noise from the NOISE92 library are used as stationary noise, emitted separately with different directions outside the camera field-of-view, and then added to the speech test signals with various signal-to-noise ratios. The environment is a quiet office. The sampling rate of signals is 16 kHz and the window length of STFT is 32 ms with overlap 16 ms.

4.3. Results

The proposed RTF estimation method was tested and compared with two existing algorithms: The non-stationarity (NS) method in [1] and the speech presence probability (SPP) method in [4]. The weight matrix in NS method is set as the identity matrix. The SPP threshold p_0 is set to 0.3, as this value achieved the best performance, and the non-recursive operator from [4] (eq. (8)) is adopted. For the proposed method, the segment length is set to 0.3s with a 50% overlap, equivalently $W = 2R = 18$. The performance metric is the average distance between the localized direction and the ground truth in the image plane (in pixels).

Table 1 shows the localization results. For WGN noise, the SPP method achieves smaller localization error than the NS method because it uses only the intervals containing speech, which decreases the error variance of RTF estimated by least-squares optimization. The proposed method obtains the best performance consistently at all SNRs, and the reasons are 1)similarly to the SPP method, we also select the segments containing speech with the proposed segment classification; 2) the proposed segmental PSD matrix subtraction accurately subtracts the noise PSD matrix, while the SPP method relies on the accuracy of noise PSD estimation; and 3) the eigenvalue decomposition for RTF estimate that is used in our method is an optimization criterion that considers all channels simultaneously. Conversely, NS and SPP methods perform least-squares optimization separately for each channel with the reference channel. Table 1 also shows the localization results for babble noise. The localization errors for babble noise are larger than WGN because of the relative non-stationarity of babble noise compared to WGN. The proposed method obtains the best performance for 5 and 10dB SNR, but SPP and the proposed method do not perform better than NS for 0dB SNR, because for babble noise, the noise PSD estimation in SPP and the proposed segment classification algorithm perform worse than for WGN.

	WGN			babble		
SIR(dB)	NS	SPP	Prop.	NS	SPP	Prop.
10	35.2	31.5	28.1	34.3	30.6	28.9
5	36.7	31.2	29.7	41.2	36.9	36.3
0	49.8	38.4	30.2	55.9	59.3	57.5
-5	107.3	64.9	41.2	-	-	-
-10	214.4	154.8	61.1	-	-	-

Table 1: Localization results for WGN noise and babble noise (in pixels, 23.3 pixels correspond to 1°).

5. CONCLUSION

We proposed an RTF identification method based on segmental PSD matrix subtraction and a classification between speech-and-noise segments and noise-only-segment based on maximum and minimum statistics of PSD. The method is shown to provide estimated RTFs that can be exploited efficiently for SSL and that competes well with other related RTF estimation methods in this context. Future work will consider the case of several sources of interest. In this case, the "desired source" in (1)−(6) would actually be the summation of multiple desired sources. If segments with negligible speech power can still be detected, the PSD matrix subtraction in (5) still makes sense but the estimation of several "individual" RTF from $\hat{\mathbf{\Phi}}(\omega)$ is not straightforward, if any feasible. However, $\hat{\mathbf{\Phi}}(\omega)$ can still be used for multiple-source application scenario, such as sound source localization based on subspace methods, which is in favor of the proposed PSD matrix subtraction method.

6. REFERENCES

[1] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. on Signal Processing*, vol. 49, pp. 1614-1626, 2001.

[2] T. G. Dvorkind and S. Gannot, "Time Difference of Arrival Estimation of Speech Source in a Noisy and Reverberant Environment," *Signal Processing*, vol. 85, No. 1, pp. 177-204, 2005.

[3] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE Trans. on Signal Processing*, vol. 4, pp. 2055-2063, 1996.

[4] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 5, pp. 451-459, 2004.

[5] K. Reindl, S. Markovich-Golan, H. Barfuss, S. Gannot, and W. Kellermann, "Geometrically constrained TRINICON-based relative transfer function estimation in underdetermined scenarios," *in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, 2013.

[6] J. Málek and Z. Koldovský, "Sparse target cancellation filters with application to semi-blind noise extraction," *in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2014.

[7] S. Markovich, S. Gannot and I. Cohen, "Multichannel Eigenspace Beamforming in a Reverberant Noisy Environment with Multiple Interfering Speech Signals," *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 17, No. 6, pp. 1071-1086, 2009.

[8] R. Martin, "Spectral subtraction based on minimum statistics," *in Proc. European Signal Processing Conference*, pp. 1182-1185, 1994.

[9] R. Martin, "Noise power spectral density estimate based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, pp. 504-512, 2001.

[10] G. Reuven, S. Gannot and I. Cohen, "Dual-source transfer-function generalized sidelobe canceller," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 711-727, 2008.

[11] M. Evans, N. Hastings and B. Peacock, "Erlang Distribution," Ch. 15 in *Statistical Distributions*, 3rd ed. New York: Wiley, pp. 84-85, 2000.

[12] A. Deleforge, V. Drouard, L. Girin and R. Horaud, "Mapping Sounds on Images Using Binaural Spectrograms", *in Proc. European Signal Processing Conference*, Lisbon, Portugal, 2014.

[13] J. Garofolo et al, TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1, Philadelphia: Linguistic Data Consortium, 1993.